Statement of Need

Machine learning (ML) offers significant potential for predictive modelling in biomedical research [@rajpurkarAIHealthMedicine2022]. In health-related contexts, predictive modelling is often used to understand the roles of prognostic factors rather than simply to classify new cases [@hastieElementsStatisticalLearning2009]. Despite its promise, there is substantial evidence that, without appropriate forethought and planning, ML models can introduce or exacerbate health inequities by making less accurate decisions for certain groups or individuals (Grote and Keeling 2022). As ML becomes increasingly embedded in healthcare systems, ensuring equitable model performance across diverse populations is essential.

The {fairmetrics} R package allows ML researchers and practitioners to evaluate group fairness of ML models via a suite of popular fairness metrics and provides estimated confidence intervals (CIs) for them through bootstrap estimation.

Fairness Criteria

Fairness of a ML model can be assessed primarily through three criteria: group fairness, individual fairness, and causal fairness. Group fairness criteria are commonly used in health and deem a model as fair if its predictions are similarly accurate or calibrated across a predefined set of groups. The groups in question are most often defined by protected attributes, such as age or race [@mehrabiSurveyBiasFairness2021].

Group fairness criteria are commonly categorized into three main types: independence, separation, and sufficiency (Barocas, Hardt, and Narayanan 2023; Berk et al. 2018). Independence requires that an ML model's predictions be statistically independent of the protected attribute. Separation demands that the model's predictions be independent of the protected attribute conditional on the true outcome class (i.e., within the positive and negative classes). Sufficiency requires that, given a model's prediction, the likelihood of the true outcome is independent of the protected attribute—aiming to equalize error rates across groups for similar prediction scores.

The {fairmetrics} package computes a range of group fairness metrics along with bootstrap-based confidence intervals. These metrics are grouped below according to the three core fairness frameworks described above.

Independence

- **Statistical Parity:** Compares the overall rate of positive predictions between groups, irrespective of the true outcome.
- **Conditional Statistical Parity:** Restricts the comparison of positive prediction rates to a specific subgroup (e.g., within a hospital unit or age bracket), offering a more context-specific fairness assessment.
- **Predictive Equality:** Compares false positive rates (FPR) between groups, ensuring that no group is disproportionately flagged as positive when the true outcome is negative.
- **Treatment Equality:** Compares the ratio of false negatives to false positives across groups, ensuring the balance of missed detections versus false alarms is consistent.

Separation

- **Accuracy Parity:** Measures whether the overall accuracy of a predictive model is equivalent across different groups.
- **Conditional Accuracy Equality:** Evaluates if predictive performance (accuracy) remains consistent across groups, conditional on the model's positive or negative decision.
- **Equalized Odds:** Simultaneously compares false negative rates (FNR) and false positive rates (FPR) between two groups defined by a binary sensitive attribute. A Bonferroni-corrected union test flags violations of equalized odds.

- **Equal Opportunity Compliance:** Focuses on disparities in false negative rates (FNR) between two groups, quantifying any difference in missed positive cases.

Sufficiency

- **Predictive Parity:** Compares positive predictive values (PPV) across groups, assessing whether the precision of positive predictions is equivalent.
- **Brier Score Parity:** Assesses whether the Brier score—the mean squared error of probabilistic predictions—is similar across groups, indicating comparable calibration.
- **Positive Class Balance:** Checks whether, among individuals whose true outcome is positive, the distribution of predicted probabilities is comparable across groups.
- **Negative Class Balance:** Checks whether, among individuals whose true outcome is negative, the distribution of predicted probabilities is comparable across groups.

Additional Features

Beyond individual metric computation, the {fairmetrics} package includes convenience functions to retrieve multiple fairness metrics (and their confidence intervals) in a single call. It also bundles example datasets based on the the MIMIC-II clinical database (J. Raffa 2016; Goldberger et al. 2000) to facilitate simple example usage of the fairness metrics with ML models.

Related Work

A similar package to the {fairmetrics} package is the {fairness} R package (Kozodoi and V. Varga 2021). The difference {fairmetrics} and {fairness} is threefold. The primary difference between the {fairmetrics} and {fairness} is that {fairmetrics} allow for the calculation of estimated confidence intervals of fairness metrics via bootstrap, which allows for more meaningful inferences about the fairness metrics calculated. Additionally, the {fairness} package has fewer dependencies and a lower memory footprint, making the for a more environment agnostic tool which can be used on modest hardware.

Licensing and Availability

The {fairmetrics} package is under the MIT liscence

ADDREF

and is available on CRAN and Github. The CRAN release can be installed with `install.packages("fairmetrics")`. For installing from Github, the {devtools} package

ADDREFERENCE

or any other R package which allows for installation of packages hosted on Github can be used (i.e. `devtools::install_github("jianhuig/fairmetrics")`).

References

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, Massachusetts: The MIT Press.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research* 50 (1): 3–44. https://doi.org/10.1177/0049124118782533.

Castelnovo, Alessandro, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. "A Clarification of the Nuances in the Fairness Metrics Landscape." *Scientific Reports* 12 (1). https://doi.org/10.1038/s41598-022-07939-1.

Fazelpour, Sina, and David Danks. 2021. "Algorithmic Bias: Senses, Sources, Solutions." *Philosophy Compass* 16 (8). https://doi.org/10.1111/phc3.12760.

Gao, Jianhui, Benson Chou, Zachary R. McCaw, Hilary Thurston, Paul Varghese, Chuan Hong, and Jessica Gronsbell. 2024. "What Is Fair? Defining Fairness in Machine Learning for Health." *arXiv.org*. https://arxiv.org/abs/2406.09307.

Grote, Thomas, and Geoff Keeling. 2022. "Enabling Fairness in Healthcare Through Machine Learning." *Ethics and Information Technology* 24 (3): 39. https://doi.org/10.1007/s10676-022-09658-7.

Kozodoi, Nikita, and Tibor V. Varga. 2021. *Fairness: Algorithmic Fairness Metrics*. https://CRAN.R-project.org/package=fairness.

Raffa, Jesse. 2016. "Clinical Data from the MIMIC-II Database for a Case Study on Indwelling Arterial Catheters (Version 1.0)." https://doi.org/10.13026/C2NC7F. https://doi.org/10.13026/C2NC7F.